United Nations ECE/CES/2017/25

Economic and Social Council

Distr.: General
11 April 2017

Original: English

Economic Commission for Europe

Conference of European Statisticians

Sixty-fifth plenary session
Geneva, 19-21 June 2017
Item 4 of the provisional agenda
The next generation of statisticians and data scientists

Statisticians or data scientists? The future of official statistics in the era of new technologies and modern data sources[1][2]

Note by the Central Bureau of Statistics of Israel

Summary

The big advancement in technology and the availability of big data pose new demands for more detailed, more accurate and timelier official statistics. The resulting technological and methodological challenges that will underlie the production of official statistics in the coming years will have major implications on the work of National Statistical Offices. The document will address issues related to the collection and management of big data for official statistics, big data privacy and protection, increasing data accessibility but maintaining privacy and confidentiality, possible use of web-panels, mode effects, social network data and integration of administrative data with surveys. This raises the question as to what extent are universities preparing students to work at modern statistical offices. This document examines several of these issues based on national and international experience.

The document is presented to the Conference of European Statisticians' seminar on "The next generation of statisticians and data scientists" for discussion.

[1] Based on the article "Methodological Issues and Challenges in the Production of Official Statistics" written by Prof. Danny Pfeffermann, National Statistician of Israel, for the 24th Annual Morris Hansen Lecture and published in the *Journal of Survey Statistics and Methodology*, December 2015. Mr. Yoel Finkel, Associate National Statistician of Israel, has edited this shortened version.

GE.17-05876(E)
1705876

I. Introduction

1. The term official statistics is used broadly, but it has never been defined formally. In this document it refers to official statistics as any set of publications by National Statistical Offices (NSOs), which are based on surveys, censuses, administrative data, or combinations of them. However, this description is rather limited because there is new extensive research underway on the use of 'big data' for the production of official statistics. Big data is generally not the result of a survey, and is usually much bigger, more dynamic and may appear in very different formats than what is traditionally perceived as administrative data. The use of big data for the production of official statistics is probably the most intriguing challenge facing NSOs. The document shall discuss this challenge in subsequent sections.

2. Official statistics are what people hear of more often than any other kind of statistics. People are exposed every month to new unemployment rates, income and poverty measures, price indexes, education achievements, health and environmental statistics, and many other related figures. For most people, official statistics is what statistics is all about. Moreover, official statistics are what policy makers use (or should use) for planning and decision-making, which affects the life of our society. When a central bank decides to change the interest rate, the decision is based on official statistics. The same is true about decisions on Government funding, building of new schools, social and health programmes and even political decisions. This being the case, it is obvious how important it is to have timely reliable official statistics for every aspect of our lives. Yet, the world is constantly changing, new advanced technologies are developing and at the same time budgets are constantly tightened.

3. The aim of this document is to discuss what may be considered as some of the major methodological challenges facing producers of official statistics and occasionally, offer ways of dealing with them. In this shortened version for the Conference of European Statisticians 2017, the following challenges are considered:

(a) Collection and management of big data for the production of official statistics;

(b) Integration of computer science for the production of official statistics from big data;

(c) Data accessibility, privacy, and confidentiality;

(d) Integration of statistics and geospatial information.

4. Having stated the immense importance of official statistics, the inevitable question that comes to mind is whether universities prepare students to work at NSOs. As the final part of this document will note, this is generally not the case. In fact, it seems that the situation has worsened in the last decade. Nowadays, only a few universities offer basic courses in official statistics, for instance on sampling or other areas. This is particularly worrying taking into account that NSOs are among the largest employers of economists and statisticians.

II. Collection and management of big data for the production of official statistics

5. The term 'big data' usually refers to large volumes of high velocity data of big value, which is complex, variable in terms of structure, sources and format, but contains also some inherent uncertainty affecting its veracity (the 5V definition, a 7V definition also exists). Typical examples are data collected on the human genome and brain, data

associated with social networks and Internet commerce, satellite readings, climate sensors, mobile phone uses, etc. The big challenges facing scientists in managing and analysing such data are discussed in numerous other publications[3]. While these excellent reports hardly mention the production of official statistics, it is obvious that NSOs cannot ignore the potential benefits of big data for statistics. Various initiatives in this direction are already taking place. For example, the Statistical Commission established in 2014 a global working group, mandated "to provide strategic vision, direction and coordination of a global programme on Big Data for official statistics, and to promote practical use of sources of Big Data for official statistics" (UN, 2014).

6. Below are some important aspects of the possible use of big data for the production of official statistics, considering computation problems and confidentiality issues in subsequent sections:

(a) Type of data: It is important to distinguish between data obtained from sensors, cameras, cellular phones, satellite images that are generally structured and accurate and relate to a particular population or area, and data obtained from social networks, e-commerce, web advertisements and alike, which are very diverse and unstructured, appears irregular, and no longer relates to a particular population. As advocated in National Research Council (2013), the structure (or lack of structure) can change over a short time, and NSOs need to be ready for that possibility. In general, data from different sources may be coded in different formats, arriving at different times with different degrees of reliability, and possibly defined differently. What is even more worrying is that some big data may suddenly cease to exist, requiring rapid modifications to the production of statistics that are based on this source. For instance, a cellular phone company may suddenly go out of business;

(b) Publication: A common feature of the data sets in the two examples outlined above and many other potential big data sets is that the data are basically available for every point in time. At present, official statistics publications are annual, monthly, or they may refer to a specific day. Three interesting questions come forth:

(i) What kind of statistics should be compiled and published? Should official publications from big data that are measured continuously be primarily in the form of (online) graphs and pictures? NSOs are already using well-developed data visualization tools, but the source data is usually much simpler;

(ii) Assuming that aggregate (average) estimates will continue to form the basis for planning and decision-making, how to transform the dynamic (continuously measurable) input data, for instance, to monthly aggregates? Should statisticians select the continuously measured data by sampling or by other, more sophisticated methods?

(iii) It seems obvious that random sampling will continue to play a major role in the era of big data, but sampling from big dynamic data will be different from sampling finite populations. This will require the development of new sampling algorithms, which not only reduce storage space but also produce manageable data sets on which algorithms can run to produce estimates, and models can be fitted, for instance the problem of sampling from social networks. Whether NSOs should use

[3] The recent report Statistics and Science (2013) includes a good summary. This report summarizes the lectures and discussions of a special workshop on the future of statistical sciences, held in London, UK, in 2013 with 100 invited participants, as part of commemorating the 'year of statistics'. Another, even more extensive (and more technical) report is the document National Research Council (2013), produced by the National Academy of Sciences of the United States.

such data for the production of official statistics is a separate issue.4 Sampling helps also to protect privacy (Section 4.1 below);

(c) Algorithmic estimation: Traditional survey sampling distinguishs between design-based estimators, model-dependent estimators, and model-assisted estimators. In the latter case the estimator is chosen based on a model, but its properties are studied under the sampling randomization distribution. With the use of big data, a new class of estimators emerges, which could be named algorithmic estimators. These estimators are the result of a computational algorithm applied to the raw data. For example, there is an ongoing request in Israel to characterize the degree of religiosity of the Jewish population, and to provide demographic and socio-economic information for the different sectors defined by this characterization. An unpublished manuscript[5], merged 12 different administrative files with the population register of Israel in early 2006. The files contained about 6 million records, and applied a complex hierarchical algorithm that assigned a religiosity score in the range of [1,3] to every person on the register. The coverage of the merged register was about 95 per cent for those between the ages of (0, 64);

(d) Measures of error: NSOs attempt to attach measures of error (uncertainty) to the published statistics, in the form of standard errors or confidence intervals. Big data are supposedly free of sampling errors (unless sampling takes place). Are measures of error still an issue in the case of big data? Should the focus be shifted to measures of bias and quality measures (measurement errors), rather than variance? How to assess the bias? Should statisticians do this by comparing the estimators with estimators obtained from a traditional survey at some point in time?[6];

(e) Bias: the potential for large bias is one of the main concerns in the use of big data for the production of official statistics. Coverage or selection bias occurs when the available data does not cover or represent correctly the whole population of interest. For example, house sale prices advertised on the Internet clearly do not represent all the sale prices in a given month (coverage bias). If the data are collected in a way that favors larger items (say, larger business), selection bias occurs. Opinions expressed in social networks are often very different from opinions held by the general public. A cheap way to deal with coverage bias, when known to exist, is to redefine the population of interest. For example, confine the population of "houses for sale" to "houses advertised "on the internet, but is this the issue of interest? In other situations, the existence of coverage or selection bias may not be known and as mentioned above, a possible way to detect and estimate the bias is by comparing estimates obtained from the big data with bias free estimates obtained from traditional surveys (as long as they continue to exist);

(f) Data linkage: NSOs not only produce and publish aggregate estimates at a national level, but very often produce estimates at much higher resolutions, as defined by age, gender, ethnicity, area of residence, type of industry, etc. However, the available big data may not contain all this information, which would require massive linkage if the missing information is available in some other sources. This, however, raises another possible limitation of big data, lack of identifiers that would allow linking different files. For example, data on purchases in supermarkets do not contain any information on the buyers, unlike the data collected in family expenditure surveys. The only identifiers that could possibly link purchases to buyers are credit card numbers, but will credit card companies release the relevant data to NSOs?.

[4] See the discussion in National Research Council (2013).

[5] Portnoi (2007) from the Israel Central Bureau of Statistics (ICBS)

[6] See National Research Council (2013) and AAPOR (2015) for more detailed discussion on possible measurement errors associated with big data

7. As a conclusion of above list, the use of big data for the production of official statistics may require new methods, for instance for data linking when extracting data from different sources. Furthermore, new methods of editing and analysis need to be developed to allow processing of the available (possibly dynamic) information with sufficient speed and accuracy, advanced visualization methods, and new methods of estimation and error assessment. These only highlight a few milestones on the way forward.

8. The next section looks at computer engineering and software developments without which the use of big data by NSOs would not be possible. Section 4 will discuss data privacy and disclosure control.

9. The document highlights the huge challenges facing computer scientists and statisticians in the use of big data. On the other hand, it would be irresponsible to ignore the potential advantages of big data for the production of official statistics in terms of timeliness (potentially in real time in some cases) and much broader versatility, coverage, and accuracy (but possibly with coverage bias). Big data are here to stay and will grow bigger and bigger. The use of big data does not require a sampling frame, questionnaires, interviews, and all other necessary ingredients underlying sample surveys. In the long run, this could result in large cost reductions. Considering that response rates in traditional surveys are constantly declining, the use of big data as an alternative or complementary source of information for official statistics seems inevitable.

III. Integration of computer science for the production of official statistics from big data

10. The very high volume and diversity of big data requires new high power hardware and software technologies for data storage and for processing and analyzing the data, which is generally not currently available at NSOs. When storing data on our laptops, we usually think in terms of gigabytes (approximately 109 bytes). Big data, however, are usually measured in terms of terabytes (1012 bytes) or petabytes (1015 bytes), but exabytes (1018 bytes) and yottabytes (1024 bytes) are also mentioned. Just to get the feeling of what it all means, Google CEO Eric Schmidt once stated that every two days we create as much information as we did from the dawn of civilization up until 2003[7]. This amounts to about five exabytes of data. The Walmart chain in the United States is said to handle one million transactions every hour, feeding a database of 2.5 petabytes, which is almost 170 times the amount of data stored in the Library of Congress in the United States. It is, therefore, not surprising that 'standard' database hardware and software tools cannot store, manage, and analyze such big data. Moreover, there is a phenomenal growth in the number of sensors and machines generating data. Examples, some of which already mentioned, include weather/pollution sensors, traffic and mobile sensors, and satellite systems.

11. A seemingly attractive solution to the need for such highly demanding computing systems is the use of cloud computing for accessing and managing very large data sets, and supports powerful infrastructure elements (storage facilities and processing power). It creates virtual machines with huge storage space and computer power. The architecture consists of arrays of virtual machines, which allow processing by segmenting into numerous parallel processes. Users (companies) can use the cloud infrastructure for their big data services without having to use their own infrastructure. In fact, cloud users do not manage the cloud infrastructure and platform where the application runs. It provides all the

[7] http://techcrunch.com/2010/08/04/schmidt-data/

necessary software, and it can also store and manage voices, which is an attractive option when thinking of the production of official statistics.

12. The cloud is likely to become increasingly important for processing big data, both for storage and access, and for analytics, even though it is still restricted at present by difficulties in transferring large data sets. It would seem therefore that the cloud can offer an attractive avenue for the use of big data by NSOs, particularly when considering that productivity can be increased when multiple users can work on the same data simultaneously. But this, in turn, highlights a major problem of data protection. In theory, data protection can be improved by data centralization, but with multiple users, and with the data distributed over a wider area or over a large number of devices, the risk of data disclosure definitely increases. Instead, private cloud (data center) installation, incorporating all local computing devices for data storage and collocated computing under centralized management should be explored. This could emerge as one of the major challenges faced by NSOs in the near future.

13. The above points only touch upon a glimpse of what big data requires in terms of computing, and what modern computing can offer. However, they emphasize how enormous the computing requirements are going to be for NSOs, if they are planning to use big data in a routine production system in terms of storage, hardware, software, computation skills, data analytics, and disclosure control.

14. Most of the traditional information infrastructures available at NSOs are not set up for these demands, requiring purchasing new proper infrastructures. It is important to mention in this respect that familiar software packages in routine use by NSOs such as SAS, SPSS and R already contain several software procedures for use with big data, but new computing skills from staff in all ranks will be required if the production of official statistics from big data is to be considered. Notwithstanding, everything said above will be treated very differently, if the cloud service would be used. In this case, the focus will be on data privacy and protection, with Government oversight and regulation. The costs involved might also be less. Either way, the potential use of big data for the production of official statistics would undoubtedly engage NSOs very extensively in coming years. At the ICBS, as in many other countries, a task force is working to look into the possibilities of using big data sets that might be available to us.

IV. Data accessibility, privacy and confidentiality

A. Preface

15. NSOs are under constant pressure from researchers, decision makers, journalists, and the general public, to release data at high resolution and if possible, to make available individual data. This, of course, is in contrast to the need to protect privacy and secure confidentiality. Without maintaining this trust, no survey can be carried out, and this obligation is emphasized in every written questionnaire or interview.

16. There are two different aspects to this issue: protecting the data from intruders, also known as ‘cyber security’, and guaranteeing that data released to people outside the NSO cannot be used to reveal private confidential data.

(a) Protecting data from intruders has to do with computer technology, a huge problem that becomes more and more dangerous, and is not restricted of course to data stored at NSOs. We are required to purchase new computer hardware and related equipment regularly to increase the protection of data from intruders. As always with this

kind of problems, in 3-4 years, statisticians may be told again that our data are no longer safe and that new expensive computing devices are needed to protect the data.

(b) The second aspect, known as statistical disclosure control, engages statisticians and computer scientists for many decades. The following reviews briefly some of the recent methods and associated quality measures in current use, with emphasis on new challenges associated with the use of big data.

B. Disclosure risks

17. Traditionally, NSOs release outputs either in the form of micro-data, mostly from social surveys, or in the form of tabular data, containing frequency counts or magnitude data typically collected in business surveys, such as total revenues. Much research has been carried out on how to quantify the disclosure risk of each of these traditional outputs under a given statistical disclosure control method, and how to assess the impact of the method used on data utility, such that the data released still contains the necessary information for research and decision making. Clearly, the further the data are protected from disclosure, the less the utility, and vice versa.

18. An emerging type of disclosure risk is inferential disclosure, which refers to learning new attributes with high probability. For example, a regression model with very high predictive power may generate inferential disclosure even for individuals who are not in the dataset. Another example of inferential disclosure is disclosure by differencing, when multiple releases are disseminated from the same data source. For example, census tables could be differenced/manipulated to reveal individual data. This kind of disclosure is best controlled by restricting to a fixed set of variables and categories, thus disallowing differencing non-nested groups of individuals.

19. A closely related concept to inferential disclosure is differential privacy, which has been widely researched by computer scientists for protecting outputs[8]. Differential privacy aims to avoid inferential disclosure by ensuring that an adversary cannot learn about the attributes of a specific target unit in the database with high probability when only one value in the database has been changed and the adversary has complete information about all the other units in the database (a 'worst case' scenario). This rather demanding definition controls disclosure resulting from differencing or highly predictive models, which becomes more problematic with increasing requests for online query systems to disseminate statistical data, compared to the use of the old hard-copy outputs. The solution offered by computer scientists to guarantee differential privacy is to add noise/perturbation to the outputs of the queries under specific parameterizations, but this of course comes at the expense of reducing data utility for inference. Consequently, other means of protecting confidentiality are constantly examined and the next sub-sections overview some of these means, concluding with a brief discussion.

C. Data protection by use of data enclaves

20. Over the last two decades, many NSOs around the world have set up on their premises research (safe) rooms, also known as data enclaves. The data enclave is a secured environment where researchers can access confidential data. The secured servers have no connection to printers or the internet, and only authorized users are allowed to access them. No data can be removed from the enclave and researchers undergo training to understand

[8] See Dinur and Nissim (2003) and Dwork, et al. (2006) for details.

the security rules. Researchers are provided with statistical software such as SAS, STATA or R, and all information flow is controlled and monitored. All outputs taken out of the data enclave are manually checked for disclosure risks such as small cell counts, residual plots, which may indicate outliers, or kernel density estimates with small band-widths.

21. The obvious disadvantages of data enclaves are the need of the researchers to travel to the NSO site and the extra burden to NSO employees to prepare the required data files and manage the enclave. Recently, some NSOs extended the concept of data enclaves to remote access via virtual data enclaves. These virtual data enclaves enable users to log on to secure servers and access the data from their personal computer, with all the activity being logged and monitored at the keystroke level. The secured data lab must be approved by the agencies and outputs are reviewed remotely by confidentiality officers before being sent back to the researchers via a secured transfer file. Obviously, the use of virtual data enclaves requires more trust by the NSO and there is less control than in the 'in house' data enclaves.

D. Statistical disclosure control for web-based applications

22. Driven by demand from policy makers and researchers for specialized tailored tables of statistical data and, in particular, census data, several NSOs have developed flexible table generating servers that allow users to define and generate their own tables. Users access the servers via the internet and define the tables from a set of accessible variables and categories.

23. There are basically two approaches of applying a statistical disclosure control method to an output table: a pre-tabular approach and a post-tabular approach. The first one applies the statistical disclosure control method to the original data so that all tables generated are deemed safe for dissemination. The latter produces first the original data and then applies a statistical disclosure control method to the table. The post-tabular approach is greatly motivated by the computer science definition of differential privacy, discussed in Section 4.2. A combination of the two approaches is also possible, although it may result in overprotection and thus reduce data utility.

24. For flexible table generation, the server has to quantify the disclosure risk in the original table, apply a statistical disclosure control method, and then reassess the disclosure risk. Clearly, the disclosure risk will depend on whether the underlying data comes from a census and the zeros are real, or whether the data is from a survey and the zeroes are random. After the table is protected, the server should also calculate the impact of applying the statistical disclosure control method on data utility, by comparing the perturbed table with the original table. Measures based on information theory can be used to assess disclosure risk and data utility in a table generating server.

25. The design of remote table generating servers typically involves many ad hoc preliminary statistical disclosure control rules that can easily be programmed within the system to exclude tables that should not be released. These statistical disclosure control rules may include limiting the number of dimensions in the table, setting minimum population thresholds such as average cell sizes or number of small cells, ensuring consistent and nested categories of variables to avoid disclosure by differencing, etc.[9]

26. Pre-tabular statistical disclosure control methods may include record swapping where attributes are swapped between two records having similar characteristics on a set of

[9] See Shlomo, Antal, and Elliott (2015) for statistical disclosure control rules and methods for remote table generating servers.

control variables. Post-tabular methods may include cell perturbation, such as random rounding, or the use of a post-randomization method, which perturbs cell counts based on a probability transition matrix. The statistical disclosure control method should ensure that sufficient statistics, such as marginal totals, are preserved and that consistency across same cells generated in different tables is maintained to avoid the possibility of recovering the statistical disclosure control method.

E. Remote analysis servers

27. A remote analysis server is an online system which accepts a query from a researcher, runs it in a secure environment on the appropriate data, and returns a confidential output, without the need for human intervention to manually check the outputs for disclosure risks. As with flexible table generators, the queries are submitted through a remote interface and researchers do not have direct access to the data. The queries may include exploratory data analysis, measures of association, regression analysis, and statistical testing. They can be run on the original data or on confidential data, and the outputs may be restricted and audited, depending on the level of protection required.[10]

F. Synthetic data

28. In recent years, there has been a move by NSOs to produce synthetic, model-based micro-data, which preserves important statistical properties of the original data. The synthetic data are stored as public-use files. The production of synthetic data becomes more and more popular since, as discussed in Section 4.3, gaining access to the real data from remote servers may be prohibited. A recent trend is to let researchers develop their research and write appropriate software code based on the synthetic data and then fit the software to the real data in a secured environment like a data enclave.

29. To produce synthetic data, a model is fitted within the agency to the original data, and the synthetic data are then sampled from the corresponding posterior distribution, similar to the theory of multiple imputation. Several samples of synthetic data may be drawn to obtain valid variance estimators[11]. Synthetic data can be combined with part of the real data so that a mixture of real and synthetic data is released[12]. Note, however, that partially synthetic datasets may still have disclosure risks that need to be checked prior to dissemination. If models used for statistical analysis are sub-models of the model used to generate data, the analysis of the synthetic data should yield valid inference, provided of course that the original model is ‘correct?.

30. For tabular data as well, there are available techniques for developing synthetic magnitude tables arising from business statistics. Controlled tabular adjustment carries out cell suppression and replaces the suppressed cells with imputed values that preserve certain statistical properties[13].

31. Should the use of synthetic data become a routine way of protecting microdata? It is the essence of statistics that analysts should use real data and not data generated from a

[10] O’Keefe and Good (2008) describe regression modeling via a remote analysis server. O’Keefe and Shlomo (2012) compare outputs based on original data and two statistical disclosure control approaches: outputs from confidential micro-data and confidential outputs obtained from the original data via a remote analysis server.

[11] See Reiter (2005) and Abowd and Vilhuber (2008) for details and discussion.

[12] See Little and Liu (2003)

[13] See Dandekar and Cox, 2002

model, although it could be argued that perturbed data are also not 'real data'. A major problem with synthetic data is that it fully depends on the model fitted to the original data, which is a subjective procedure, and the model may not capture all the relationships between variables, particularly within sub-populations. It is also difficult to reproduce possible abnormalities in the data. For example, the original data may contain outlying observations for particular sets of units (say, businesses). If the synthetic data is supposed to resemble the original data, then it should also contain outlying observations of similar magnitude and behaviour, for similar sets of units. Is the confidentiality of the data still preserved even if the outlying observations are somewhat changed? Finally, what about big data? Are we going to generate many sets of big synthetic data, thus amplifying the problem of storage and management of big data by several factors?

G. Discussion

32. There is an obvious conflict between the demand from researchers for more detailed data and the responsibility of NSOs to protect the confidentiality of respondents. This conflict generates an enormous need for statistical disclosure control methods that address the dual tasks of guaranteeing confidentiality with very high probability and preserving the utility of the data released to researchers. This has led to close cooperation between computer scientists who have developed formal definitions of disclosure risk, particularly for inferential disclosure, and statisticians developing statistical disclosure control methods that guarantee confidentiality. The use of these methods implies that researchers have to cope with perturbed data when carrying out statistical analysis, requiring therefore increased knowledge on statistical inference under measurement errors. There is an obvious need to further explore perturbative methods of statistical disclosure control that preserve data utility, thus allowing for consistent and unbiased estimation of statistical models.

33. What about big data? The problem of data dissemination becomes larger requiring even more tightened cooperation with computer analysts. To begin with, we will need to deal with much larger volumes of high dimensional complex data, with many more variables and categories than in traditional surveys. Restricting to samples selected from the big data is one way to proceed, not only to reduce the size of the data but also as a mean of preserving data confidentiality.

34. There are public policy and ethical problems with the dissemination of big data by NSOs, which may require special legislation. When carrying out a survey or a census, there is a clear commitment to preserve the confidentiality of the data. No such commitment is given in conjunction with big data assembled from sensors, cell phone companies or social networks. Will companies assembling the big data be required to transfer the data to the NSO? Will the public agree with the transfer of private data to researchers, and then possible dissemination, even if applying statistical disclosure control? Furthermore, with potential access to several data sets and the possibility to link them, there is an increased potential for breaking confidentiality. This is particularly true for data sets tracking human activities, for example, identifying the same person in multiple social media sources.

V. Integration of statistics and geospatial information

35. The use of Geographic Information Systems (GIS) enables the addition of a spatial dimension to the data collected and hence the possibility to get new insights into the data.[14]

[14] The Morris Hansen Lecture by Michael Goodchild in 2006 discusses this theme in great depth (Goodchild, 2007).

A well-known example is the poverty maps produced by the World Bank and other organizations, where every geographic region on the map is coloured, with different colours representing different levels of poverty. Another example is a road accidents map, in which case each section on the road is coloured based on the number of road accidents. The prominent advantage of this kind of hot spot maps is that they not only show at a glance which geographic locations (regions, road sections…) require extra attention, but they also show spatial similarities between neighbouring locations if they exist. To put it differently, it transforms discrete estimates into a continuum.

36. The use of GIS has many other important benefits:

- It enhances the design of sample surveys by defining the borders of strata, sampling cells etc. It also provides all the required information for area sampling;
- It enables efficient allocation of sampling quotas for interviewers, and the construction of optimal navigation routes to arrive at the sampled locations;
- The use of GIS permits tracing phenomena such as changes over time in socio economic conditions of individuals or households residing in given areas, and relating them to geographic measures such as distance from a big city, commuting possibilities and movements to other areas;
- GIS enhances data resolution very significantly, enabling the exploration and formation of new groupings (clusters) that were not known before.

37. The rapid advancement of technology opens the way for the collection of new (big) data in very high resolution, and the use of GIS will enable relating this data to very detailed geographical locations.

38. So far, the document has noted that the future statistician will not only need to be trained in classical statistical theory but will need training in computer science and cyber-security. Three additional important issues for NSOs, with implications for future statisticians, include: possible use of web panels for the production of official statistics; handling of mode effects in mixed-mode surveys; and the combination of administrative data with small area estimation. These issues are discussed at great length in the original document and the advanced statistical methods presented in the article were eliminated from this shortened version.

VI. Are universities preparing students to work at National Statistical Offices?

A. Preface

39. The introduction raised the question of whether universities prepare students to work at NSOs. This is generally not the case, but the need for more and better training has actually been raised in several other forums and some positive actions are being taken. The starting point for this discussion is that no one questions the importance of NSOs and other similar organizations, and that NSOs are among the largest employers of statisticians and economists.

40. This section will discuss the following three central topics in the work of NSOs and consider to what extent they are taught at universities.

1. Survey sampling

41. There is obviously no need to elaborate in this article on the importance of survey sampling for the work of NSOs. It is impossible to design a survey, edit (clean) the data correcting for outliers and nonresponse, and produce proper estimates and estimates of error, without good knowledge of the theory of survey sampling. The following is the curriculum of a one year course on Design and Analysis of Sample Surveys, offered by Harvard University (3 hours a week):

Methods for design and analysis of sample surveys; the toolkit of sample design features and their use in optimal design strategies; sampling weights and variance estimation methods, including resampling methods; brief overview of non-statistical aspects of survey methodology such as survey administration and questionnaire.

2. Seasonal adjustment (SA) and trend estimation

42. Time series of socio-economic data are used for studying trends and detecting changes (turning points) in socio-economic activity. This learning process, however, is impossible when the observed time series encompasses not only the trend-cycle component of interest, but also seasonal movements, trading day effects, and moving holidays and irregular influences. These additional components have to be estimated and removed from the observed series if the trend is to be studied. A first estimate of the trend is obtained by subtracting the seasonal effects, known as seasonal adjustment. Several model-dependent and nonparametric procedures have been proposed in the literature for seasonal adjustment and trend estimation and are in routine use. It is common practice among many NSOs to publish the seasonally adjusted or the trend series along with the original (observed) series of interest. The difference between the two component series is that the seasonally adjusted series contains also the irregular terms, which are smoothed out for estimating the trend. The trend series is smoother, but it may hide important turning points towards the end of the series. Conversely, the seasonally adjusted series may exhibit false turning points.

3. National accounts

43. A fundamental part of the work of NSOs is devoted to the production of high quality economic statistics. Examples include national accounts, balance of payments, government finance statistics, price statistics, international trade statistics, satellite accounts (energy, welfare, education, etc.), and environmental-economic accounts. The system of national accounts (SNA) is one of the most important systems of series and is produced under strict international standards. It provides users and decision makers with a comprehensive understanding of the economic activity and its evolution. In Israel, like in many other countries, analysts, decision makers, the media and the general public are awaiting very eagerly the new NA estimates, every time that they are released. One would assume therefore that such an important foundation of macroeconomic statistics would be taught at every department of economics of every university. Is this the case?

B. Answer to question

44. Are universities teaching students the basic knowledge needed for working in these three important areas? In order to answer this question, we browsed the curriculum of the undergraduate and postgraduate courses in statistics and economics at the top 25 universities in the world, as ranked by the Shanghai Academic Ranking of World Universities, http://www.shanghairanking.com/ (see the Annex). Here is what we learnt:

45. Only 11 out of the top 25 universities offer some form of an introductory course on survey sampling.

46. While almost all the universities offer one or more courses on *Time series Analysis and Forecasting*, there are no courses that devote a significant portion of the time to seasonal adjustment or trend estimation. In fact, only three time series courses mention seasonality in the description.

47. Specialized courses on national accounts are even rarer. At best, macroeconomics courses occasionally mention national accounts, mostly with reference to the gross domestic product (GDP). It seems only the International Monetary Fund and its affiliates provide comprehensive courses on national accounts, in addition to a few Master Programmes in Official Statistics (see below).

48. Being faced with these findings, the natural question to ask is what is actually the role of public or private universities? Are they at all supposed to teach and train their students to work at work places, or should universities focus solely on research, and educate new generations of researchers? This debate goes back many centuries and is obviously beyond the scope of this document. However, the following points should be considered:

(a) The three topics mentioned, and many other topics underlying the work of NSOs require a good deal of knowledge in theory of statistics or economics. Survey sampling, seasonal adjustment and trend estimation involve advanced topics in theoretical statistics with very important applications, so teaching courses on these topics is not in conflict with the view that universities should concentrate on research. Some of the leading mathematical statisticians in the world are involved in research on small area estimation, another very important production of NSOs. The following list includes a few examples of topics underlying the work of NSOs based on classical statistical theory: Survey sampling methods, fitting models to complex survey data, linkage procedures, statistical disclosure control, signal extraction of ARIMA models, estimation of the mean squared error (MSE) of seasonally adjusted estimates, the use of resampling methods for bias reduction and variance estimation;

(b) It may be that the reason for the shortage of courses on topics related to the work of NSOs is the lack of expert researchers to teach them. Therefore, students are not exposed to survey sampling and other important problems underlying the work of NSOs during their studies, and hence they do not consider these problems in their PhD thesis or later for their academic research;

(c) Statistics have changed very dramatically in the last decade or so, with what is known as ‘classical statistics’ giving way to new sophisticated computer intensive methods of analysing big data and alike. Survey sampling and time series analysis are also not the same as they used to be, and it is clear that courses on such topics need to be restructured, accounting for modern developments in these and other areas, thus hopefully making them also more attractive;

(d) After all, quite a few universities in several countries emphasize survey sampling in their teaching and research activities. Moreover, there are several university masters' programmes in official statistics. Two well-known programmes are the Joint Programme in Survey Methodology in the United States and Westat and the Master's Programme in Official Statistics at the University of Southampton in the United Kingdom. The first is a collaboration between the University of Maryland and the University of Michigan, and the latter is sponsored by the United Kingdom Office for National Statistics (ONS). The National Institute for Statistics and Economics Studies (INSEE) in France, and the Brazilian Institute of Geography and Statistics (IBGE) have schools (colleges) of statistics with special programmes for Bachelor and Master's degrees in official statistics.

(e) Finally, the European Union has recently decided to establish a European Master Programme in Official Statistics (EMOS), and contracts with European NSOs and

Universities are already tendered. The European Master in Official Statistics (EMOS) is a network of Master programmes providing post-graduate education in the area of official statistics at the European level. EMOS is a joint project of universities and data producers in Europe. After two calls for interest, the network comprises over 20 programmes in 14 countries.

(f) EMOS was set up to strengthen the collaboration within academia and producers of official statistics and help develop professionals able to work with European official data at different levels in the fast-changing production system of the 21st century. The EMOS Master degree bases on learning outcomes that familiarize the graduates with the system of official statistics, production models, statistical methods and dissemination.

(g) Universities offering EMOS Master Degrees collaborate actively with the NSOs to reduce the gap between theory and practice. Therefore, there is a growing recognition of the importance of the production of official statistics by academic institutions, and this recognition may spread to other universities.

VII. References[15] [English only]

AAPOR (2010). *Report on online survey panels*. http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.html?

AAPOR (2015). *Report on big data.* American Association for Public Opinion Research. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

Abowd, J.M., and Vilhuber, L. (2008). *How protective are synthetic data?* In: PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin). Springer LNCS 5262, 239-246.

Cavallo, A., and Rigobon, R. (2010). *The Billion Prices Project@MIT*. (http://bpp.mit.edu).

Cavallo, A. (2012). *Online vs official price indexes: measuring Argentina's inflation.* Journal of Monetary Economics, 1-14.

Chaudhuri, S., Handcock, M.S. and Rendall, M.S. (2010). *A conditional empirical likelihood approach to combine sampling design and population level information.* Technical report No. 3/2010, National University of Singapore, Singapore, 117546.

Couper, M.P. (2000). *Web surveys, a review of issues and approaches*. Public Opinion Quarterly 64, 464-494.

Couper, M.P. (2008). *Designing Effective Web Surveys.* Cambridge University Press.

Daas P.J.H., Puts M.J. Buelens B. and van den Hurk, P.A.M. (2013*). Big Data and official statistics. Proceedings of the NTTS*, Euro Stat, Brussels. http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf

Dandekar, R.A., and Cox L.H. (2002). *Synthetic tabular data: an alternative to complementary cell suppression.* Manuscript, Energy Information Administration, U. S. Department of Energy.

De Leeuw, E. (2005*). To mix or not to mix? Data collection modes in Surveys*. Journal of Official Statistics, 21, 1-23.

Dillman, D.A., and Christian, L. (2005). *Survey mode as a source of instability in response across surveys.* Field Methods, 17, 30-52.

Dinur, I., and Nissim, K. (2003). *Revealing Information While Preserving Privacy.* PODS 2003, 202-210.

Dwork, C., McSherry, F. Nissim, K. and Smith, A. (2006). *Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC* (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.

Fay, R.E., and Herriot, R. (1979). *Estimates of income for small places: an application of James–Stein procedures to census data.* Journal of the American Statistical Association, 74, 269–77.

Feder, M., and Pfeffermann, D. (2015). *Statistical inference under non-ignorable sampling and nonresponse- an empirical likelihood approach.* Southampton Statistical Sciences Research Institute, http://eprints.soton.ac.uk/id/eprint/378245

[15] References listed relate to those from the published article in the ***Journal of Survey Statistics and Methodology***, December 2015.

Goodchild, M.F. (2007). *The Morris Hansen Lecture 2006: Statistical perspectives on social science.* Journal of Official Statistics, 23, 1–15.

Hartley, H.O., and Rao, J.N.K. (1968). *A new estimation theory for sample surveys*. Biometrika, 55, 547-557.

Lee, J., and Berger, J.O. (2001). *Semiparametric Bayesian Analysis of Selection Models*. Journal of the American Statistical Association, 96, 1397-1409.

Lee, S. (2006). *Propensity score adjustment as a weighting scheme for volunteer panel web surveys.* Journal of Official Statistics, 22, 329-349.

Lee, S., and Valliant, R. (2009). *Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment*. Sociological Methods & Research, 37. 319-343.

Little, R.J.A., and Liu, F. (2003). *Selective multiple imputation of keys for statistical disclosure control in microdata*. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6.

National Research Council (2013). *Frontiers in Massive Data Analysis*. Washington D.C.: The National Academies Press. (http://www.nap.edu)

New Zealand (2012). *Using cellphone data to measure population movements*. (http://www.stats.govt.nz/tools_and_services/earthquake-info-portal/using-cellphone-data-report.aspx).

Nirel, R. and Glickman, H. (2009*). Sample surveys and censuses.* In: Handbook of Statistics 29A. Sample Surveys: Design, Methods and Application Eds., D. Pfeffermann and C.R. Rao. Amsterdam: North Holland, 539-565.

O'Keefe, C.M. and Good, N. (2008*). A remote analysis server – What Does Regression Output Look Like?* In PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 270-283.

O'Keefe, C.M. and Shlomo, N. (2012). *Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data*. Transactions on Data Privacy, 5, 403-432.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd edition). New York: Cambridge University Press.

Pfeffermann, D. (2013*). New important developments in small area estimation*. Statistical Science, 28, 40-68.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998). *Parametric distributions of complex survey data under informative probability sampling*. Statistica Sinica, 8, 1087-1114.

Pfeffermann, D. and Landsman, V. (2011*). Are private schools better than public schools?* Appraisal for Ireland by methods for observational studies. The Annals of Applied Statistics, 5, 1726–1751.

Pfeffermann, D., Moura, F. A. S. and Nascimento-Silva, P.L. (2006). *Multilevel modeling under informative sampling.* Biometrika, 93, 943-959.

Pfeffermann, D. and Sikov A. (2011). *Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information.* Journal of Official Statistics, 27, 181-209.

Pfeffermann, D. and Sverchkov, M. (1999). *Parametric and semi-parametric estimation of regression models fitted to survey data.* Sankhya, 61, 166-186.

Pfeffermann, D. and Sverchkov, M. (2003). *Fitting generalized linear models under informative probability sampling.* In: Analysis of Survey Data, eds. R. L. Chambers and C. J. Skinner, New York: Wiley, pp. 175-195.

Pfeffermann., D. and Sverchkov, M. (2007). *Small area estimation under informative probability sampling of areas and within the selected areas*. Journal of the American Statistical Association, 102, 1427-1439.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ.MR1953089

Reiter, J.P. (2005). *Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical study.* Journal of the Royal Statistical Society, A, 168, 185-205.

Rivers, D. (2007). *Sampling for web surveys*. Joint Statistical Meeting, Proceedings of the Section on Survey Research Methods, Salt Lake City, UT, USA.

Rosenbaum, P.R. and Rubin, D.B. (1983). *The central role of the propensity score in observational studies for treatment effects.* Biometrika, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984). *Reducing bias in observational studies using subclassification on the Propensity score.* Journal of the American Statistical Association, 79, 516-524.

Rotnitzky, A. and Robins, J. (1997). *Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response*. Statistics in Medicine, 16, 81-102.

Shlomo, N., Antal, L. and Elliot, M. (2015). *Measuring disclosure risk and data utility for flexible table generators*. Journal of Official Statistics, 31, 305–324.

Smith T.M.F. (1994). *Sample surveys 1975-1990; an age of reconciliation*? International Statistical Review, 62, 5-19.

Statistics and Science (2013). *A report of the London workshop on the future of the statistical sciences.* http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

Sverchkov, M., and Pfeffermann, D. (2004). *Prediction of finite population totals based on the sample distribution.* Survey Methodology, 30, 79-92.

UN (2014). *Report of the global working group on big data for official statistics.* United Nations, E/CN.3/2015/4. http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData-E.pdf

Waksberg, J. and Goldfield, E. D. (1996). Morris Howard Hansen, 1920-1990. *A biographical memoir.* National Academy of Sciences, Washington D.C., U.S.A.

Vaccari, C. (2014). *Big Data in Official statistics. School of advanced studies*, University of Camerino, Italy. https://www.academia.edu/7571682/PhD).

Vannieuwenhuyze, J.T.A; Loosveldt, G and Molenberghs, G. (2014). *Evaluating mode effects in mixed-mode survey data using covariate adjustment models*. Journal of Official Statistics, 30, 1-21.